(54) Title: SYSTEM AND METHOD FOR MODIFYING SPEECH SIGNALS
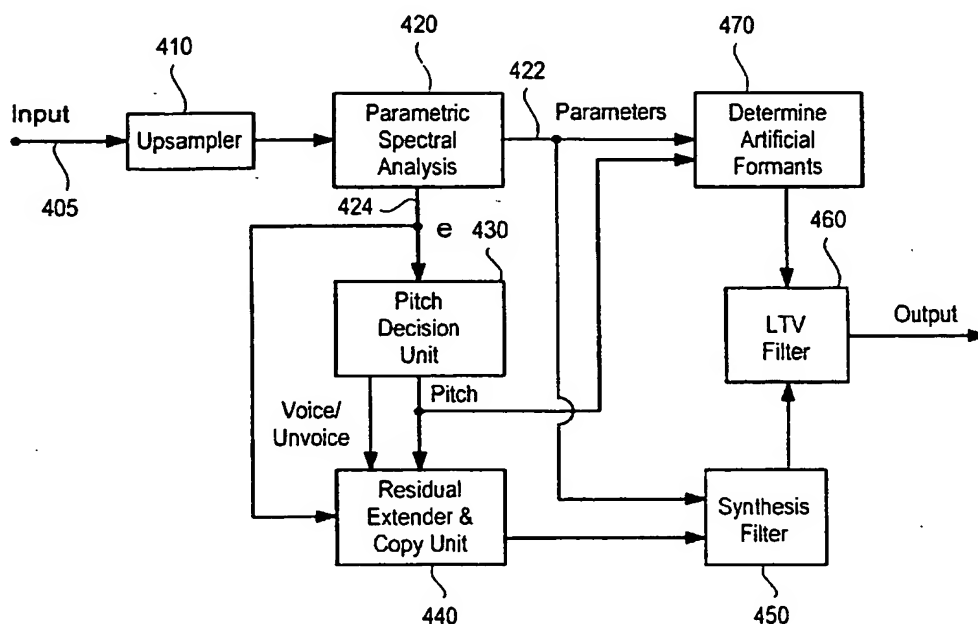
(57) Abstract: A system and method for speech signal enhancement upsamples a narrowband speech signal at a receiver to generate a wideband speech signal. The lower frequency range of the wideband speech signal is reproduced using the received narrowband speech signal. The received narrowband speech signal is analyzed to determine its formants and pitch information. The upper frequency range of the wideband speech signal is synthesized using information derived from the received narrowband speech signal.

before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SYSTEM AND METHOD FOR MODIFYING SPEECH SIGNALS

BACKGROUND

The present invention relates to techniques for transmitting voice information in communication networks, and more particularly to techniques for enhancing narrowband speech signals at a receiver.

In the transmission of voice signals, there is a trade off between network capacity (*i.e.*, the number of calls transmitted) and the quality of the speech signal on those calls. Most telephone systems in use today encode and transmit speech signals in the narrow frequency band between about 300 Hz and 3.4 kHz with a sampling rate of 8 kHz, in accordance with the Nyquist theorem. Since human speech contains frequencies between about 50 Hz and 13 kHz, sampling human speech at an 8kHz rate and transmitting the narrow frequency range of approximately 300 Hz to 3.4 kHz necessarily omits information in speech signal. Accordingly, telephone systems necessarily degrade the quality of voice signals.

Various methods of extending the bandwidth of speech signals transmitted in telephone systems have been developed. The methods can be divided into two categories. The first category includes systems that extend the bandwidth of the speech signal transmitted across the entire telephone system to accommodate a broader range of frequencies produced by human speech. These systems impose additional bandwidth requirements throughout the network, and therefore are costly to implement.

A second category includes systems that use mathematical algorithms to manipulate narrowband speech signals used by existing phone systems. Representative examples include speech coding algorithms that compress wideband speech signals at a transmitter, such that the wideband signal may be transmitted across an existing narrowband connection. The wideband signal must then be de-compressed at a receiver. These methods can be expensive to implement since the structure of the existing systems need to be changed.

-2-

Other techniques implement a "codebook" approach. A codebook is used to translate from the narrowband speech signal to the new wideband speech signal. Often the translation from narrowband to wideband is based on two models: one for narrowband speech analysis and one for wideband speech synthesis. The codebook is

5   trained on speech data to "learn" the diversity of most speech sounds (phonemes). When using the codebook, narrowband speech is modeled and the codebook entry that represents a minimum distance to the narrowband model is searched. The chosen model is converted to its wideband equivalent, which is used for synthesizing the wideband speech. One drawback associated with codebooks is that they need significant training.

10   Another method is commonly referred to as spectral folding. Spectral folding techniques are based on the principle that content in the lower frequency band may be folded into the upper band. Normally the narrowband signal is re-sampled at a higher sampling rate to introduce aliasing in the upper frequency band. The upper band is then shaped with a low-pass filter, and the wideband signal is created. These methods are

15   simple and effective, but they often introduce high frequency distortion that makes the speech sound metallic.

Accordingly, there is a need in the art for additional systems and methods for transmitting narrowband speech signals. Further, there is a need in the art for systems and methods for processing narrowband speech signals at a receiver to simulate wideband

20   speech signals.

SUMMARY

The present invention addresses these and other needs by adding synthetic information to a narrowband speech signal received at a receiver. Preferably, the speech

25   signal is spilt into a vocal tract model and an excitation signal. One or more resonance frequencies may be added to the vocal tract model, thereby synthesizing an extra formant in the speech signal. Additionally, a new synthetic excitation signal may be added to the original excitation signal in the frequency range to be synthesized. The speech may then be synthesized to obtain a wideband speech signal. Advantageously, methods of the

-3-

invention are of relatively low computational complexity, and do not introduce significant distortion into the speech signal.

In one aspect, the present invention provides a method for processing a speech signal. The method comprises the steps of: analyzing a received, narrowband signal to
5   determine synthetic upper band content; reproducing a lower band of the speech signal using the received, narrowband signal; and combining the reproduced lower band with the determined, synthetic upper band to produce a wideband speech signal having a synthesized component.

According to further aspects of the invention, the step of analyzing further
10   comprises the steps of: performing a spectral analysis on the received narrowband signal to determine parameters associated with a speech model and a residual error signal; determining a pitch associated with the residual error signal; identifying peaks associated with the received, narrowband signal; and copying information from the received, narrowband signal into an upper frequency band based on at least one of the determined
15   pitch and the identified peaks to provide the synthetic upper band content.

According to further aspects of the invention, a predetermined frequency range of the wideband signal may be selectively boosted. The wideband signal may also be converted to an analog format and amplified.

In accordance with another aspect, the invention provides a system for processing
20   a speech signal. The system comprises means for analyzing a received, narrowband signal to determine synthetic upper band content; means for reproducing a lower band of the speech signal using the received, narrowband signal; and means for combining the reproduced lower band with the determined, synthetic upper band to produce a wideband speech signal having a synthesized component.

25   According to further aspects of the system, the means for analyzing a received, narrowband signal to determine synthetic upper band content comprises: a parametric spectral analysis module for analyzing the formant structure of the narrowband signal and generating parameters descriptive of the narrow band voice signal and an error signal; a pitch decision module for determining the pitch of the sound segment represented by the
30   narrowband signal; and a residual extender and copy module for processing information

-4-

derived from the narrowband voice signal and generating a synthetic upper band signal component.

According to additional aspects of the invention, the residual extender and copy module comprises a Fast Fourier Transform module for converting the error signal from the parametric spectral analysis module into the frequency domain; a peak detector for identifying the harmonic frequencies of the error signal; and a copy module for copying the peaks identified by the peak detector into the upper frequency range.

In yet another aspect, the invention provides a system for processing a narrowband speech signal at a receiver. The system includes an upsampler that receives the narrowband speech signal and increases the sampling frequency to generate an output signal having an increased frequency spectrum; a parametric spectral analysis module that receives the output signal from the upsampler and analyzes the output signal to generate parameters associated with a speech model and a residual error signal; a pitch decision module that receives the residual error signal from the parametric spectral analysis module and generates a pitch signal that represents the pitch of the speech signal and an indicator signal that indicates whether the speech signal represents voiced speech or unvoiced speech; and a residual extender and copy module that receives and processes the residual error signal and the pitch signal to generate a synthetic upper band signal component.

BRIEF DESCRIPTION OF THE DRAWINGS

The objects and advantages of the invention will be understood by reading the following detailed description in conjunction with the drawings, in which:

Fig. 1 is a schematic depiction illustrating the functions of a receiver in accordance with aspects of the invention;

Fig. 2 illustrates a representative spectrum of voiced speech and the coarse structure of the formants;

Fig. 3 illustrates a representative spectrogram;

-5-

Fig. 4 is a block diagram illustrating one exemplary embodiment of a system and method for adding synthetic information to a narrowband speech signal in accordance with the present invention;

Fig. 5 is a block diagram illustrating an exemplary residual extender and copy circuit depicted in Fig. 4;

Fig. 6 is a block diagram illustrating a second exemplary embodiment of a system and method for adding synthetic information to a narrowband speech signal in accordance with the present invention;

Fig. 7 is a block diagram illustrating an exemplary residual extender and copy circuit depicted in Fig. 6;

Fig. 8 is a block diagram illustrating a third exemplary embodiment of a system and method for adding synthetic information to a narrowband speech signal in accordance with the present invention;

Fig. 9 is a block diagram illustrating an exemplary residual modifier in accordance with the present invention;

Fig. 10 is a graph illustrating a short-time autocorrelation function of a speech sample that represents a voiced sound;

Fig. 11 is a graph illustrating an average magnitude difference function of a speech sample that represents a voiced sound;

Fig. 12 is a block diagram illustrating that an AR model transfer function may be separated into two transfer functions;

Fig. 13 is a graph illustrating the coarse structure of a speech signal before and after adding a synthetic formant to the speech signal;

Fig. 14 is a graph illustrating the coarse structure of a speech signal before and after adding a synthetic formant to the speech signal; and

Fig. 15 is a graph illustrating the frequency response curves of AR models having different parameters on a speech signal.

-6-

DETAILED DESCRIPTION

The present invention provides improvements to speech signal processing that may be implemented at a receiver. According to one aspect of the invention, frequencies of the speech signal in the upper frequency region are synthesized using information in the lower frequency regions of the received speech signal. The invention makes advantageous use of the fact that speech signals have harmonic content, which can be extrapolated into the higher frequency region.

The present invention may be used in traditional wireline (*i.e.*, fixed) telephone systems or in wireless (*i.e.*, mobile) telephone systems. Because most existing wireless phone systems are digital, the present invention may be readily implemented in mobile communication terminals (*e.g.*, mobile phones or other communication devices). Fig. 1 provides a schematic depiction of the functions performed by a communication terminal acting as a receiver in accordance with aspects of the present invention. An encoded speech signal is received by the antenna 110 and receiver 120 of a mobile phone, is decoded by a channel decoder 130 and a vocoder 140. The digital signal from vocoder 140 is directed to a bandwidth extension module 150, which synthesizes missing frequencies of the speech signal (*e.g.*, information in the upper frequency region) based on information in the received speech signal. The enhanced signal may be transmitted to a D/A converter 160, which converts the digital signal to an analog signal that may be directed to speaker 170. Since the speech signal is already digital, the sampling is already performed in the transmitting mobile phone. It will be appreciated, however, that the present invention is not limited to wireless networks; it can generally be used in all bidirectional speech communication.

Speech Production

By way of background, speech is produced by neuromuscular signals from the brain that control the vocal system. The different sounds produced by the vocal system are called phonemes, which are combined to form words and/or phrases. Every language has its own set of phonemes, and some phonemes exist in more than one language.

-7-

Speech-sounds may be classified into two main categories: voiced sounds and unvoiced sounds. Voiced sounds are produced when quasi-periodic bursts of air are released by the glottis, which is the opening between the vocal cords. These bursts of air excite the vocal tract, creating a voiced sound (i.e., a short "a" (ä) in "car"). By contrast,

5   unvoiced sounds are created when a steady flow of air is forced through a constraint in the vocal tract. This constraint is often near the mouth, causing the air to become turbulent and generating a noise-like sound (i.e., as "sh" in "she"). Of course, there are sounds which have characteristics of both voiced sounds and unvoiced sounds.

There are a number of different features of interest to speech modeling techniques.

10   One such feature is the formant frequencies, which depend on the shape of the vocal tract. The source of excitation to the vocal tract is also an interesting parameter.

Fig. 2 illustrates the spectrum of voiced speech sampled at a 16 kHz sampling frequency. The coarse structure is illustrated by the dashed line 210. The three first formants are shown by the arrows.

15   Formants are the resonance frequencies of the vocal tract. They shape the coarse structure of the speech frequency spectrum. Formants vary depending on characteristics of the speaker's vocal tract, i.e., if it is long (typical for male), or short (typical for female). When the shape of the vocal tract changes, the resonance frequencies also change in frequency, bandwidth, and amplitude. Formants change shape continuously

20   during phonemes, but abrupt changes occur at transitions from a voiced sound to an unvoiced sound. The three formants with lowest resonance frequencies are important for sampling the produced speech sound. However, including additional formants (e.g., the 4th and 5th formants) enhances the quality of the speech signal. Due to the low sampling rate (i.e., 8kHz) implemented in narrowband transmission systems, the higher-frequency

25   formants are omitted from the encoded speech signal, which results in a lower quality speech signal. The formants are often denoted with F_k where k is the number of the formant.

There are two types of excitation to the vocal tract: impulse excitation and noise excitation. Impulse excitation and noise excitation may occur at the same time to create a

30   mixed excitation.

-8-

Bursts of air originating from the glottis are the foundation of impulse excitation. Glottal pulses are dependent on the sound pronounced and the tension of the vocal cords. The frequency of glottal pulses is referred to as the fundamental frequency, often denoted F_o. The period between two successive bursts is the pitch-period and it ranges from approximately 1.25 ms to 20 ms for speech, which corresponds to a frequency range between 50 Hz to 800 Hz. The pitch exists only when the vocal cords vibrate and a voiced sound (or mixed excitation sound) is produced.

Different sounds are produced depending on the shape of the vocal tract. The fundamental frequency F_o is gender dependent, and is typically lower for male speakers than female speakers. The pitch can be observed in the frequency-domain as the fine structure of the spectrum. In a spectrogram, which plots signal energy (typically represented by a color intensity) as a function of time and frequency, the pitch can be observed as the thin horizontal lines, as depicted in Fig. 3. This structure represents the pitch frequency and it's higher order harmonics originating from the fundamental frequency.

When unvoiced sounds are produced the source of excitation represents noise. Noise is generated by a steady flow of air passing through a constriction in the vocal tract, often in the oral cavity. As the flow of air passes the constriction it becomes turbulent, and a noise sound is created. Depending on the type of phoneme produced the constriction is located at different places. The fine structure of the spectrum differs from a voiced sound by the absence of the almost equally spaced peaks.

Exemplary Speech Signal Enhancement Circuits

Fig. 4 illustrates an exemplary embodiment of a system and method for adding synthetic information to a narrowband speech signal in accordance with the present invention. Synthetic information can be added to a narrowband speech signal to expand the reproduced frequency band, thereby providing improved reproduced perceived speech quality. Referring to Fig. 4, an input voice or speech signal 405 received by a receiver, (*e.g.*, a mobile phone), is first upsampled by upsampler 410 to increase the sampling frequency of the received signal. In a preferred embodiment, upsampler 410 may

upsample the received signal by a factor of two (2), but it will be appreciated that other upsampling factors may be applied.

The upsampled signal is analyzed by a parametric spectral analysis module 420 to determine the formant structure of the received speech signal. The particular type of

5   analysis performed by parametric spectral analysis unit 420 may vary. In one embodiment, an autoregressive (AR) model may be used to estimate model parameters as described below. Alternatively, a sinusoidal model may be employed in parametric spectral analysis unit 420 as described, for example, in the article entitled "Speech Enhancement Using State-based Estimation and Sinusoidal Modeling" authored by

10   Deisher and Spanias, the disclosure of which is incorporated here by reference. In either case, the parametric spectral analysis unit 420 outputs parameters, (i.e., values associated with the particular model employed therein) descriptive of the received voice signal, as well as an error signal (e) 424, which represents the prediction error associated with the evaluation of the received voice signal by parametric spectral analysis unit 420.

15   The error signal (e) 424 is used by pitch decision unit 430 to estimate the pitch of the received voice signal. Pitch decision unit 430 can, for example, determine the pitch based upon a distance between transients in the error signal These transients are the result of pulses produced by the glottis when producing voiced sounds. Pitch decision module 430 also determines whether the speech content of the received signal represents a voiced

20   sound or an unvoiced sound, and generates a signal indicative thereof. The decision made by the pitch decision unit 430 regarding the characteristic of the received signal as being a voiced sound or an unvoiced sound may be a binary decision or a soft decision indicating a relative probability of a voiced signal or an un-voiced signal.

The pitch information and a signal indicative of whether the received signal is a

25   voiced sound or an unvoiced sound are output from the pitch decision unit 430 to a residual extender and copy unit 440. As described below with respect to Fig. 5, the residual extender and copy unit 440 extracts information from the received narrow band voice signal, (e.g., in the range of 0 to 4 kHz) and uses the extracted information to populate a higher frequency range, (e.g., 4 kHz-8 kHz). The results are then forwarded to

30   a synthesis filter 450, which synthesizes the lower frequency range based on the

-10-

parameters output from parametric spectral analysis unit 420 and the upper frequency range based on the output of the residual extender and copy unit 440. The synthesis filter 450 can, for example, be an inverse of the filter used for the AR model. Alternatively, synthesis filter 450 can be based on a sinusoidal model.

5   A portion of the frequency range of interest may be further boosted by providing the output of the synthesis filter 450 to a linear time variant (LTV) filter 460. In one exemplary embodiment, LTV filter 460 may be an infinite impulse response (IIR) filter. Although other types of filters may be employed, IIR filters having distinct poles are particularly suited for modeling the voice tract. The LTV filter 460 may be adapted based

10   upon a determination regarding where the artificial formant (or formants) should be disposed within the synthesized speech signal. This determination is made by determination unit 470 based on the pitch of the received voice signal as well as the parameters output from parametric spectral analysis unit 420 based on a linear or nonlinear combination of these values, or based upon values stored in a lookup table and

15   indexed based on the derived speech model parameters and determined pitch.

          Fig. 5 depicts an exemplary embodiment of residual extender and copy unit 440. Therein, the residual error signal (e) 424 from parametric spectral analysis unit 420 is input to a Fast Fourier Transform (FFT) module 510. FFT unit 510 transforms the error signal into the frequency domain for operation by copy unit 530. Copy unit 530, under

20   control of peak detector 520, selects information from the residual error signal (e) 424 which can be used to populate at least a portion of an excitation signal. In one embodiment, peak detector 520 may identify the peaks or harmonics in the residual error signal (e) 424 of the narrowband voice signal. The peaks may be copied into the upper frequency band by copy module 530. Alternatively, peak detector 520 can identify a

25   subset of the number of peaks, (e.g., the first peak), found in the narrowband voice signal and use the pitch period identified by pitch decision unit 430 to calculate the location of the additional peaks to be copied by copy unit 530. The signal that indicates whether the sampled narrowband signal is a voiced sound or an unvoiced sound also is provided to peak detector 520 since peak detection and copying are replaced by artificial unvoiced

30   upper band speech content when the speech segment represents an unvoiced sound.

-11-

Unvoiced speech content is generated by speech content unit 540. Artificial unvoiced upper band speech content can be created in a number of different ways. For example, a linear regression dependent on the speech parameters and pitch can be performed to provide artificial unvoiced upper band speech content. As an alternative, an associated memory module may include a look-up table that provides artificial upper band unvoiced speech content corresponding to input values associated with the speech parameters derived from the model and the determined pitch. The copied peak information from the residual error signal and the artificial unvoiced upper band speech content are input to combination module 560. Combination unit 560 permits the outputs of copy unit 530 and artificial unvoiced upper band speech content unit 540 to be weighted and summed together prior to being converted back into the time domain by FFT unit 570. The weight values can be adjusted by gain control unit 550. Gain control module 550 determines the flatness of the input spectrum, and uses this information and pitch information from pitch decision module 430, regulates the gains associated with the combination unit 120. Gain control unit 550 also receives the signal indicating whether the speech segment represents a voiced sound or an unvoiced sound as part of the weighting algorithm. As described above, this signal may be binary or "soft" information that provides a probability of the received signal segment being processed being either a voiced sound or an unvoiced sound.

Fig. 6 illustrates another exemplary embodiment of a system and method for adding a synthetic voice formant to an upper frequency range of a received signal. The embodiment depicted in Fig. 6 is similar to the embodiment depicted in Fig. 4, except that the residual extender and copy module 640 provides an output which is based only on information copied from the narrowband portion of the received signal. An exemplary embodiment of this residual extender and copy module 640 is illustrated as Fig. 7, and is described below. If the pitch decision unit 630 determines that a particular segment of interest represents an unvoiced sound, it controls switch 635 to select the residual error (e) signal directly for input to synthesis filter 650. By contrast, if pitch decision module 630 determines that a voice signal is present, then switch 635 is controlled to be connected to the output of residual extender and copy unit 640 such that the upper frequency content is

-12-

determined thereby. A boost filter 660 operates on the output of synthesis filter 650 to increase the gain in a predetermined portion of the desired sampling frequency. For example, boost filter 660 can be designed to increase the gain the band from 2 kHz to 8 kHz. By simulating the reproduction of various synthetic voice formants as described

5   herein, the filter pole pairs can be optimized, for example, in the vicinity of a radius of 0.85 and an angle of 0.58 π.

Fig. 7 provides an example of a residual extender and copy unit 640 employed in the exemplary embodiment of Fig. 6. Therein, the residual error signal (e) is once again transformed into the frequency domain by FFT unit 710. Peak detector 720 identifies

10   peaks associated with the frequency domain version of the residual error signal (e), which are then copied by copy module 730 and transformed by into the time domain by FFT module 740. As in the exemplary embodiment of Fig. 5 peak detector 620 can detect each of the peaks independently, or a subset of the peaks, and can calculate the remaining peaks based upon the determined pitch. As will be apparent to those skilled in the art, this

15   particular implementation of the residual extender and copy module is somewhat simplified when compared with the implementation in Fig. 5 since it does not attempt to synthesize unvoiced sounds in the upper band speech content.

Fig. 8 is a schematic depiction of another exemplary embodiment of a system and method for adding a synthetic voice formant to an upper frequency range of a received

20   signal in accordance with the present invention. A narrowband speech signal, denoted by $x(n)$ is directed to an upsampler 810 to obtain a new signal $s(n)$ having an increased sampling frequency of, e.g., 16 kHz. It will be noted that n is the sample number. The upsampled signal $s(n)$ is directed to a Segmentation module 820 that collects the set of samples comprising the signal $s(n)$ into a vector (or buffer).

25   The formant structure can be estimated using, for example, an AR model. The model parameters , a_k, can be estimated using, for example, a linear prediction algorithm. A linear prediction module 840 receives the upsampled signal $s(n)$ and the sample vector produced by Segmentation module 820 as inputs, and calculates the predictor polynomial a_k, as described in detail below. A Linear Predictive Coding (LPC) module 830 employs

30   the inverse polynomial to predict the signal $s(n)$ resulting in a residual signal $e(n)$, the

-13-

prediction error. The original signal is recreated by exciting the AR model with the residual signal $e(n)$.

The signal is also extended into the upper part of the frequency band. To excite the extended signal, the residual signal $e(n)$ is extended by the residual modifier module 860, and is directed to a synthesizer module 870. In addition, a new formant module 850 estimates the positions of the formants in the higher frequency range, and forwards this information to the synthesizer module 870. The synthesizer module 870 uses the LPC parameters, the extended residual signal, and the extended model information supplied by new formant module 850 to create the wide band speech signal, which is output from the system.

Fig. 9 illustrates a system for extending the residual signal into the upper frequency region, which may correspond to residual modifier module 860 depicted in Fig. 8. The residual signal $e_i(n)$ is directed to a pitch estimation module 910, which determines the pitch based upon, e.g., a distance between the transients in the error signal and generates a signal 912 representative thereof. Pitch estimation module 910 also determines whether the speech content of the received signal is a voiced sound or an unvoiced sound, and generates a signal 914 indicative thereof. The decision made by the pitch estimation module 910 regarding the characteristic of the received signal as being a voiced sound or an unvoiced sound may be a binary decision or a soft decision indicating a relative probability that the signal represents a voiced sound, or an unvoiced sound. Residual signal $e_i(n)$ is also directed to a first FFT module 920 to be transformed into the frequency domain, and to a switch 950. The output of first FFT module 920 is directed to a modifier module 930 that modifies the signal to a wideband format. The output of modifier module 930 is directed to an inverse FFT (IFFT) module 940, the output of which is directed to switch 950.

If the pitch estimation module 910 determines that a particular segment of interest represents an unvoiced sound, then it controls switch 950 to select the residual error (e) signal directly for input to synthesizer 870. By contrast, if pitch estimation module 910 determines that the segment represents a voiced sound, then switch 950 is controlled to be connected to the output of modifier module 930 and IFFT module 940, such that the

-14-

upper frequency content is determined thereby. The output from switch 950 may be directed, *e.g.*, to synthesizer 870 for further processing.

The systems described in Fig. 8 and Fig. 9 may be used to implement two methods of populating the upper frequency band. In a first method, modifier 930 creates harmonic peaks in the upper frequency band by copying parts of the lower band residual signal to the higher band. The harmonic peaks may be aligned by finding the first harmonic peak in the spectrum that reaches above the mean of the spectrum and last peak within the frequency bins corresponding to the telephone frequency band. The section between the first and last peak may be copied to the position of the last peak. This results in equally spaced peaks in the upper frequency-band. Although this method may not make the peaks reach to the end of the spectrum (8kHz), the technique can be repeated until the end of the spectrum has been reached.

The result of this process is depicted in Fig. 13, which reflects substantially equally spaced peaks in the upper frequency band. Since there is only one synthetic formant added in the vicinity of 4.6 kHz, there is no formant model that can be excited by harmonics over approximately 6 kHz. This method does not create any artifacts in the final synthetic speech. Depending on the amount of noise added in the calculation of the AR model, the extended part of the spectrum may need to be weighted with a function that decays with increasing frequency.

In the second method, modifier module 930 uses the pitch period to place the new harmonic peaks in the correct position in the. By using the estimated pitch-period it is possible to calculate the position of the harmonics in the upper frequency band, since the harmonics are assumed to be multiples of the fundamental frequency. This method makes it possible to create the peaks corresponding to the higher order harmonics in the upper frequency band.

In the Global System for Mobile communications (GSM) telephone system, the transmissions between the mobile phone and the base station are done in blocks of samples. In GSM the blocks consists of 160 samples corresponding to 20 ms of speech. The block size in GSM assumes that speech is a quasi-stationary signal. The present invention may be adapted to fit the GSM sample structure, and therefore use the same

-15-

block size. One block of samples is called a frame. After upsampling, the frame length
will be 320 samples and is denoted with L.

The AR Model of Speech Production

5   One way of modeling speech signals is to assume that the signals have been
created from a source of white noise that has passed through a filter. If the filter consists
of only poles, the process is called an autoregressive process. This process can be
described by the following difference equation when assuming short time stationarity.

10

$$s_i(n) = \sum_{k=1}^{p} a_{ik} s_i(n-k) + w_i(n) \tag{1}$$

where $w_i(n)$ is white noise with unit variance, $s_i(n)$ is the output of the process and p is the
model order. The $s_i(n-k)$ is the old output values of the process and a_{ik} is the

15   corresponding filter coefficient. The subscript i is used to indicate that the algorithm is
based on processing time-varying blocks of data where i is the number of the block. The
model assumes that the signal is stationary during in the current block, i. The
corresponding system-function in the z-domain may be represented as:

20

$$H_i(z) = \frac{1}{1 - \sum_{k=1}^{p} a_{ik} z^{-k}} = \frac{1}{A_i(Z)} \tag{2}$$

where $H_i(z)$ is the transfer function of the system and $A_i(z)$ is called the predictor. The
system consists of only poles and does not fully model the speech, but it has been shown

25   that when approximating the vocal apparatus as a loss-less concatenation of tubes the
transfer function will match the AR model. The inverse of the system function for the AR
model, an all-zeros function is

$$\frac{1}{H_i(z)} = 1 + \sum_{k=1}^{p} a_{ik} z^{-k} = A_i(Z) \tag{3}$$

30

-16-

which is called the prediction filter. This is the one-step prediction of $s_i(n+1)$ from the last $p+1$ values of $[s_i(n), ..., s_i(n-p+1)]$. The predicted signal called $\hat{s}_i(n)$ subtracted from the signal $s_i(n)$ yields the prediction error $e_i(n)$, which is sometimes called the residual. Even though this approximation is incomplete, it provides valuable information about the speech signal. The nasal cavity and the nostrils have been omitted in the model. If the order of the AR model is chosen sufficiently high, then the AR model will provide a useful approximation of the speech signal. Narrowband speech signals may be modeled with an order of eight (8).

The AR model can be used to model the speech signal on a short term basis, *i.e.*, typical segments of 10-30 ms of duration, where the speech signal is assumed to be stationary. The AR model estimates an all-pole filter that has an impulse response, $\hat{s}_i(n)$, that approximates the speech signal, $s_i(n)$. The impulse response, $\hat{s}_i(n)$, is the inverse z-transform of the system function H(z). The error, $e(n)$, between the model and the speech signal can then be defined as

$$e_i(n) = s_i(n) - \hat{s}_i(n) - s_i(n) - \sum_{k=1}^{p} a_{ik}(i) s_i(n-k) \qquad (4)$$

There are several methods for finding the coefficients, a_{ik}, of the AR model. The autocorrelation method yields the coefficients that minimize

$$\varepsilon(i) = \sum_{n=0}^{L+p-1} |e_i(n)|^2 \qquad (5)$$

where L is the length of the data. The summation starts at zero and ends at $L+p-1$. This assumes that the data is zero outside the L available data and is accomplished by multiplying $s_i(n)$ with a rectangular window. Minimizing the error function results in solving a set of linear equations

-17-

$$
\begin{bmatrix}
r_{\overline{si}}(0) & r_{\overline{si}}(1) & \cdots & r_{\overline{si}}(p-1) \\
r_{\overline{si}}(1) & r_{\overline{si}}(0) & \cdots & r_{\overline{si}}(p-2) \\
\vdots & \vdots & \ddots & \\
r_{\overline{si}}(p-1) & r_{\overline{si}}(p-2) & \cdots & r_{\overline{si}}(0)
\end{bmatrix}
\begin{bmatrix}
a_{i1} \\
a_{i2} \\
\vdots \\
a_{ip}
\end{bmatrix}
=
\begin{bmatrix}
r_{\overline{si}}(1) \\
r_{\overline{si}}(2) \\
\vdots \\
r_{\overline{si}}(p)
\end{bmatrix}
\qquad (6)
$$

where $r_{si}(k)$ represents the autocorrelation of the windowed data (n) and a_{ik} is the coefficients of the AR model.

Equation 6 can be solved in several different ways, one method is the Levinson-Durbin recursion, which is based upon the fact that the coefficient matrix is Toeplitz. A matrix is Toeplitz if the elements in each diagonal have the same value. This method is fast and yields both the filter coefficients, a_{ik}, and the reflection coefficients. The reflection coefficients are used when the AR model is realized with a lattice structure. When implementing a filter in the fixed-point environment, which often is the case in mobile phones, insensitivity to quantization of the filter-coefficients should be considered. The lattice structure is insensitive to these effects and is therefore more suitable than the direct form implementation. A more efficient method for finding the reflection-coefficients is Schur's recursion, which yields only the reflection-coefficients.

Pitch Determination

Before the pitch-period can be estimated the nature of the speech segment must be determined. The predictor described below results in a residual signal. Analyzing the residual speech signal can reveal whether the speech segment represents a voiced sound or an unvoiced sound. If the speech segment represents an unvoiced sound, then the residual signal should resemble noise. By contrast, if the residual signal consists of a train of impulses, then it is likely to represent a voiced sound. This classification can be done in many ways, and since the pitch-period also needs to be determined, a method that can estimate both at the same time is preferable. One such method is based on the short-time normalized auto-correlation function of the residual signal defined as

-18-

$$R_{ie}(l) = \frac{1}{R_{ie}(0)} \sum_{n=0}^{L-1-l} e_i(n) e_i(n+l) \tag{7}$$

where n is the sample number in the frame with index i, and l is the lag. The speech signal is classified as voiced sound when the maximum value of $R_{ie}(l)$ is within the pitch range and above a threshold. The pitch range for speech is 50-800 Hz, which corresponds to l in the range of 20-320 samples. Fig. 10 shows a short-time auto-correlation function of a voiced frame. A peak is clearly visible around lag 72. Peaks are also visible at multiples of the fundamental frequency.

Another algorithm suitable for analyzing the residual signal is the average magnitude difference function (AMDF). This method has a relatively low computational complexity. This method also uses the residual signal. The definition of the AMDF is

$$AMDF_i(l) = \frac{1}{L} \sum_{n=0}^{L-l} |e_i(n) - e_i(n-l)| \tag{8}$$

This function has a local minimum at the lag corresponding to the pitch-period. The frame is classified as voiced sound when the value of the local minimum is below a variable threshold. This method needs at least a data-length of two pitch-periods to estimate the pitch-period. Fig. 11 shows a plot of the AMDF function for a voiced frame, several local minima can be seen. The pitch period is about 72 samples which means that the fundamental frequency is 222 Hz when the sampling frequency is 16 kHz.

Adding a Synthetic Formant

Different methods to add synthetic resonance frequencies have been evaluated. All these methods model the synthetic formant with a filter.

The AR model has a transfer function of the form

$$H_i(z) = \frac{1}{1 - \sum_{k=1}^{p} a_{ik} z^{-k}} \tag{9}$$

-19-

which can be reformulated as

$$H_i(z) = \frac{1}{(1 - \sum_{k=1}^{p-2} a_{ik}^l z^{-k})} - \frac{1}{1 + a_{i(p-1)}^l z^{-1} + a_{if}^l z^{-2}} = H_{i1}(z) \cdot H_{i2}(z) \tag{10}$$

where a_{ik}^l represents the two new AR model coefficients. As illustrated in Fig. 12, one filter can be divided into two filters. $H_{i1}(z)$ represents the AR model calculated from the current speech segment and $H_{i2}(z)$ represent the new synthetic formant filter.

5   In one method, the synthetic formant(s) are represented by a complex conjugate pole pair. The transfer function $H_{i2}(z)$ may then be defined by the following equation:

$$H_{i2}(z) = \frac{b_0}{1 - 2v\cos(\omega_s) + v^2} \tag{11}$$

where v is the radius and ω_5 is the angle of the pole. The parameter b_0 may be used to set
10   the basic level of amplification of the filter. The basic level of amplification may be set to 1 to avoid influencing the signal at low frequencies. This can be achieved by setting b_o equal to the sum of the coefficients in $H_{i2}(z)$ denominator. A synthetic formant can be placed at a radius of 0.85 and an angle of 0,58π. Parameter b_0 will then be 2.1453. If this synthetic formant is added to the AR model estimated on the narrowband speech signal,
15   then the resulting transfer function will not have a prominent synthetic formant peak. Instead, the transfer function will lift the frequencies in the range 2.0-3.4 kHz. The reason that the synthetic formant is not prominent is because of large magnitude level differences in the AR model, typically 60-80 dB. Enhancing the modified signal so that the formants reach an accurate magnitude level decreases the formant bandwidth and
20   amplifies the upper frequencies in the lower band by a few dB. This is illustrated in Fig. 13, in which dashed line 1310 represents the coarse spectral structure before adding a synthetic formant. Solid line 1320 represents the spectral structure after adding a synthetic formant, which generates a small peak at approximately 4.6 kHz.

-20-

Thus, a formant filter that uses one complex conjugate pole pair renders it difficult to make the formant filter behave like an ordinary formant. If high-pass filtered white noise is added to the speech signal prior to the calculation of the AR model parameters, then the AR model will model the noise and the speech signal. If the order of the AR model is kept unchanged (*e.g.*, order eight), some of the formants may be estimated poorly. When the order of the AR model is increased so that it can model the noise in the upper band without interfering with the modeling of the lower band speech signal, a better AR model is achieved. This will make the synthetic formant appear more like an ordinary formant. This is illustrated in Fig. 14, in which dashed line 1410 represents the coarse spectral structure before adding a synthetic formant. Solid line 1420 represents the spectral structure after adding a synthetic formant, which generates a peak at approximately 4.6 kHz.

Fig. 15 illustrates the difference between the AR model calculated with and without the added noise to the speech signal. Referring to Fig. 15, the solid line 1510 represents an AR model of the narrowband speech signal, determined to the fourteenth order. Dashed line 1520 represents an AR model of the narrowband speech signal, determined to the fourteenth order, and supplemented with high pass filtered noise. Dotted line 1530 represents an AR model of the narrowband speech signal determined to the eighth order.

Another way to solve the problem is to use a more complex formant filter. The filter can be constructed of several complex conjugate pole pairs and zeros. Using a more complicated synthetic formant filter increases the difficulty of controlling the radius of the poles in the filter and fulfilling other demands on the filter, such as obtaining unity gain at low frequencies.

To control the radius of the poles of the synthetic formant filter, the filter should be kept simple. A linear dependency between the existing lower frequency formants and the radius of the new synthetic formant may be assumed according to

$$v_1 \alpha_1 + v_2 \alpha_2 + v_3 \alpha_3 + v_4 \alpha_4 = v_{\omega s} \qquad (12)$$

-21-

where v_1, v_2, v_3 and v_4 are the radius of the formants in the AR model from the narrowband speech signal. Parameters α_m, $m = 1,2,3,4$ are the linear coefficients. Parameter $v_{\omega 5}$ is the radius of the synthetic fifth formant of the AR model of the wideband speech signal. If several AR models are used then equation 12 can be expressed as

$$\begin{vmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & r_{k3} & r_{k4} \end{vmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} r_{15w} \\ r_{25w} \\ \vdots \\ r_{k5w} \end{bmatrix} \tag{13}$$

where v are the formant radius and the first index denote the AR model number, the second index denotes formant number and the third index w in the rightmost vector denotes the estimated formant from the wideband speech signal, and k is the number of AR models. This system of equations is overdetermined and the least square solution may be calculated with the help of the pseudoinverse.

The solution obtained was then used to calculate the radius of the new synthetic formant as

$$\hat{v}_{15} = r_{11}\alpha_1 + r_{12}\alpha_2 + r_{13}\alpha_3 + r_{14}\alpha_4 \tag{14}$$

where $\hat{v}_{15}$, is the new synthetic formant radius and the α-parameters are the solution for the equation system 13.

The present invention is described above with reference to particular embodiments, and it will be readily apparent to those skilled in the art that it is possible to embody the invention in forms other than those described above. The particular embodiments described above are merely illustrative and should not be considered restrictive in any way. The scope of the invention is determined given by the following claims, and all variations and equivalents that fall within the range of the claims are intended to be embraced therein.

-22-

What is claimed is:

1.   A method for processing a speech signal, comprising the steps of:
analyzing a received, narrowband signal to determine synthetic upper band

5   content;

reproducing a lower band of the speech signal using the received,
narrowband signal; and

combining the reproduced lower band with the determined, synthetic upper
band to produce a wideband speech signal having a synthesized component.

10

2.   The method of claim 1, wherein the received, narrowband signal provides
information content in the range of about 0-4 kHz and the synthetic upper band content is
in the range of about 4-8kHz.

15   3.   The method of claim 1, wherein the step of analyzing further comprises the
steps of:

performing a spectral analysis on the received narrowband signal to
determine parameters associated with a speech model and a residual error signal;

determining a pitch associated with the residual error signal;

20   identifying peaks associated with the received, narrowband signal; and

copying information from the received, narrowband signal into an upper
frequency band based on at least one of the determined pitch and the identified peaks to
provide the synthetic upper band content.

25   4.   The method of claim 3, wherein the step of performing a spectral analysis
employs an AR-predictor.

5.   The method of claim 4, wherein the step of performing a spectral analysis
employs a sinusoidal model.

30

-23-

6.   The method of claim 1, further comprising the step of selectively boosting a predetermined frequency range of the wideband signal.

7.   The method of claim 1, further comprising the step of converting the wideband signal to an analog format.

8.   The method of claim 7, further comprising the step of amplifying the wideband signal.

9.   A system for processing a speech signal, comprising:
         means for analyzing a received, narrowband signal to determine synthetic upper band content;
         means for reproducing a lower band of the speech signal using the received, narrowband signal; and
         means for combining the reproduced lower band with the determined, synthetic upper band to produce a wideband speech signal having a synthesized component.

10.   A system according to claim 9, wherein the means for analyzing a received, narrowband signal to determine synthetic upper band content comprises:
         a parametric spectral analysis module for analyzing the formant structure of the narrowband signal and generating parameters descriptive of the narrow band voice signal and an error signal;
         a pitch decision module for determining the pitch of the sound segment represented by the narrowband signal; and
         a residual extender and copy module for processing information derived from the narrowband voice signal and generating a synthetic upper band signal component.

-24-

11.   A system according to claim 10, wherein the residual extender and copy module comprises:

a fast Fourier transform module for converting the error signal from the parametric spectral analysis module into the frequency domain;

5   a peak detector for identifying the harmonic frequencies of the error signal; and

a copy module for copying the peaks identified by the peak detector into the upper frequency range.

10   12.   A system according to claim 11, wherein the residual extender and copy module further comprises:

a module for generating artificial unvoiced speech content.

13.   A system according to claim 12, wherein the residual extender and copy module further comprises:

15

a combiner for combining an output signal from the copy module and an output from the module fro generating artificial unvoiced speech content.

14.   A system according to claim 13, wherein the residual extender and copy module further comprises:

20

a gain control module for weighting the input signals in the combiner.

15.   A system according to claim 13, wherein the residual extender and copy module further comprises:

25   a fast Fourier transform module for converting the error signal from the parametric spectral analysis module from the frequency domain into the time domain.

16.   A system according to claim 9, wherein the means for reproducing a lower band of the speech signal using the received, narrowband signal comprises:

-25-

a parametric spectral analysis module for analyzing the formant structure of the narrowband signal and generating parameters descriptive of the narrowband voice signal and an error signal; and

a synthesis filter.

17.   A system for processing a narrowband speech signal at a receiver, comprising:

an upsampler that receives the narrowband speech signal and increases the sampling frequency to generate an output signal having an increased frequency spectrum;

a parametric spectral analysis module that receives the output signal from the upsampler and analyzes the output signal to generate parameters associated with a speech model and a residual error signal;

a pitch decision module that receives the residual error signal from the parametric spectral analysis module and generates a pitch signal that represents the pitch of the speech signal and an indicator signal that indicates whether the speech signal represents voiced speech or unvoiced speech;

a residual extender and copy module that receives and processes the residual error signal and the pitch signal to generate a synthetic upper band signal component.

18.   A system according to claim 17, further comprising:

a synthesis filter that receives parameters from the parametric spectral analysis module and information derived from the residual error signal, and generates a wideband signal that corresponds to the narrowband speech signal.

19.   A system according to claim 17, wherein the indicator signal from the pitch decision module controls a switch connected to an input to the synthesis filter, such that if the indicator signal indicates that the speech signal represents voiced speech, then the input to the synthesis filter is connected to the output of the residual extender and copy module, and if the indicator signal indicates that the speech signal represents

-26-

unvoiced speech, then the input to the synthesis filter is connected to the residual error signal output from the parametric spectral analysis module.
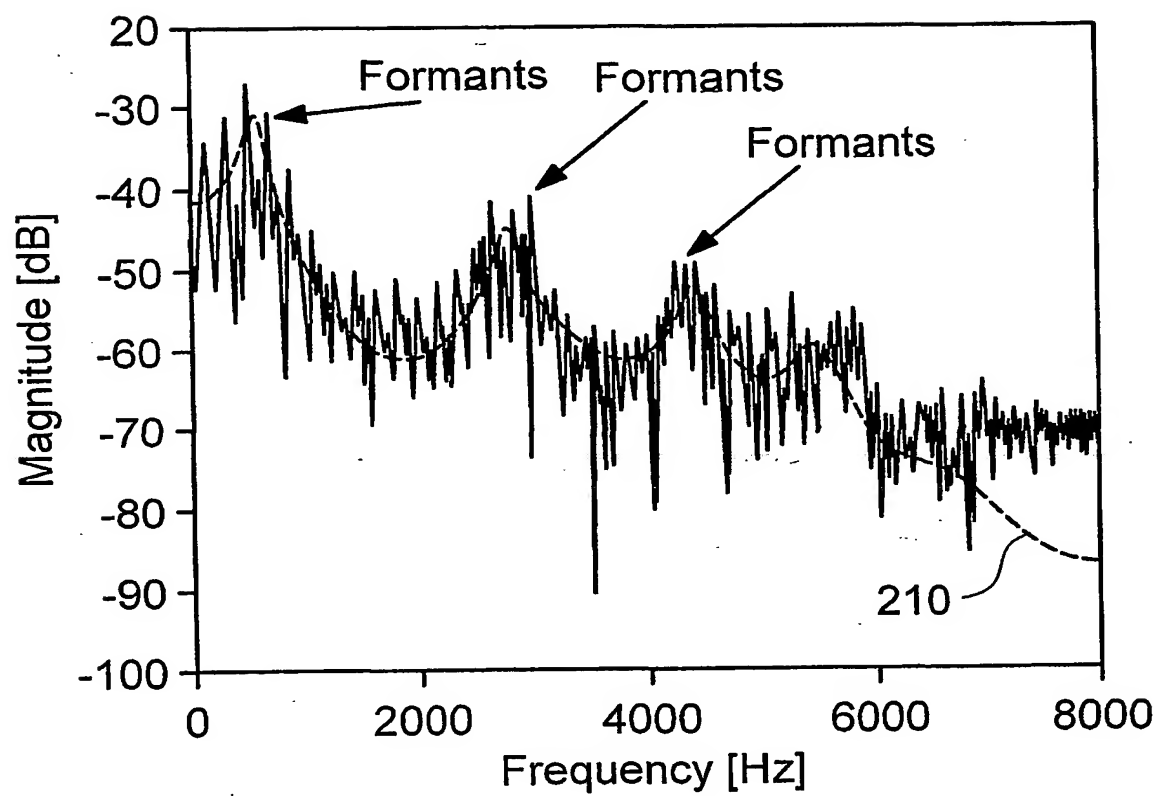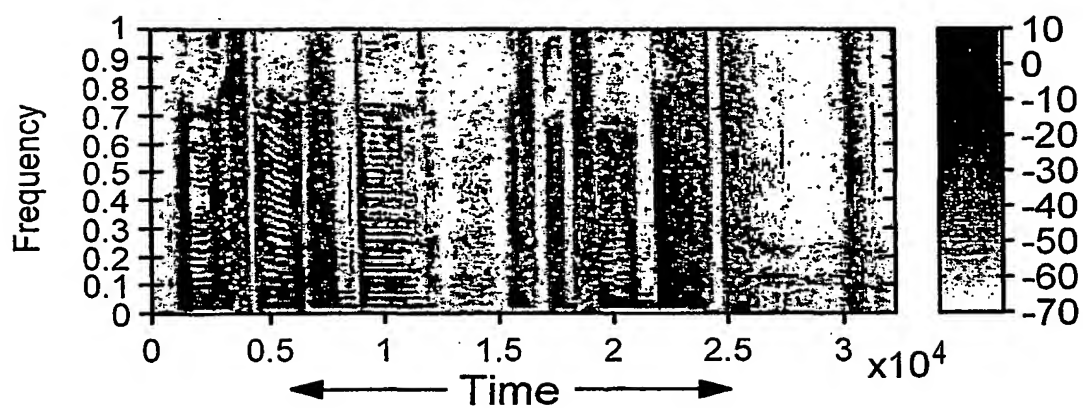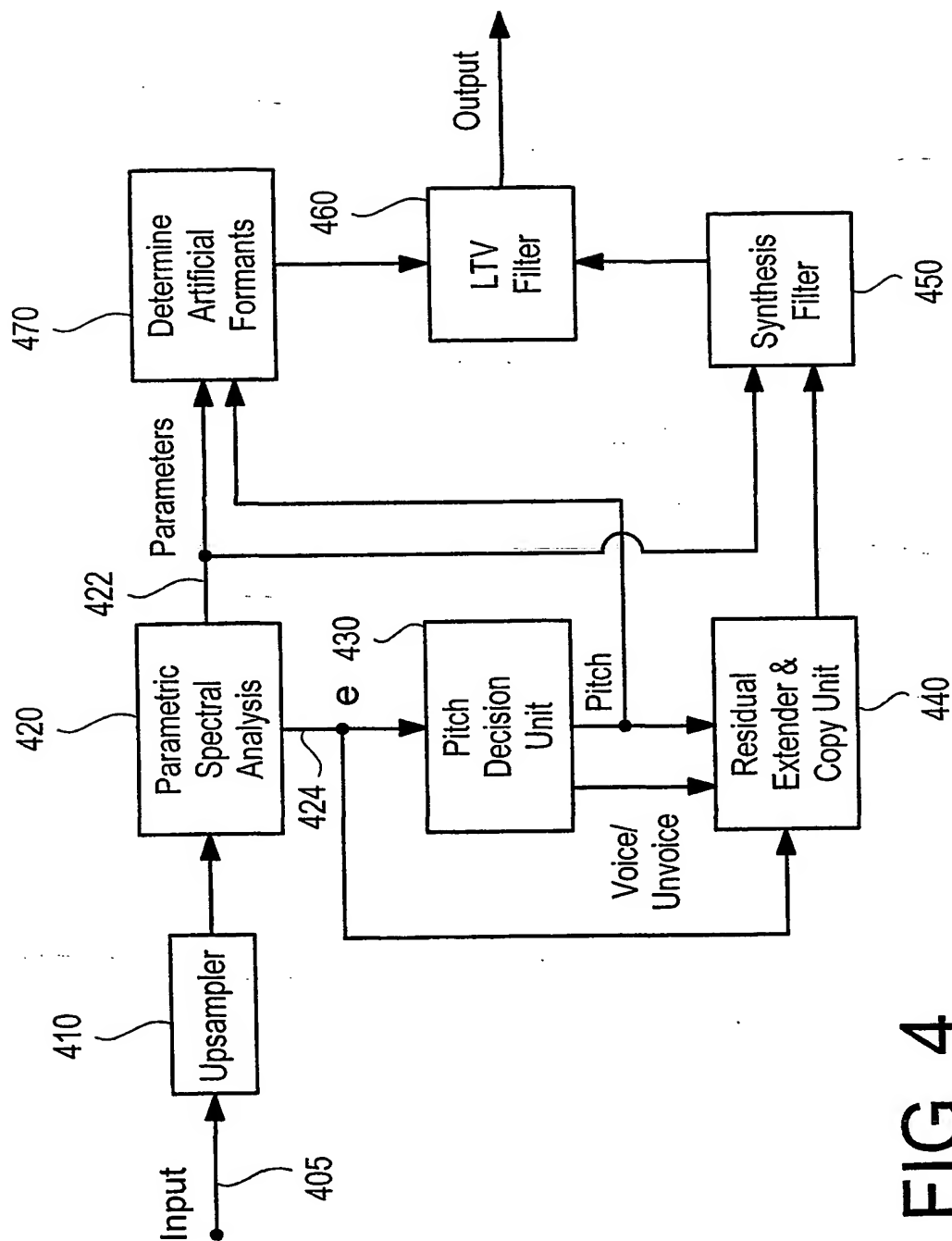
1/15

FIG. 1
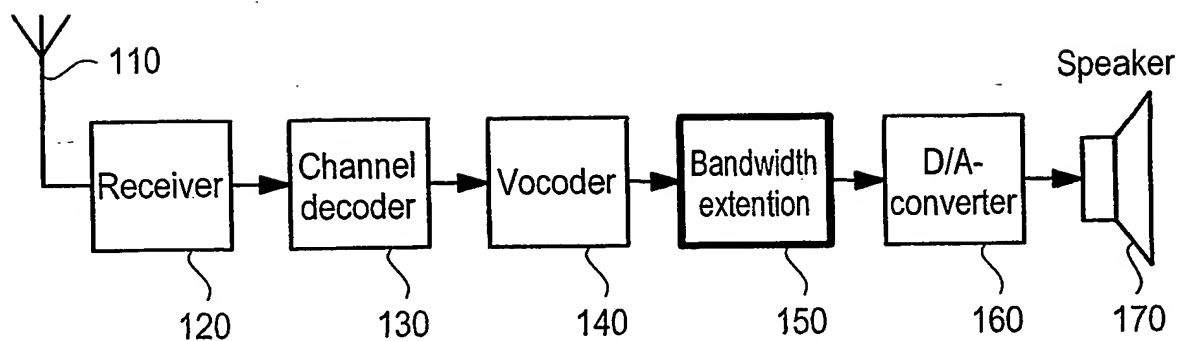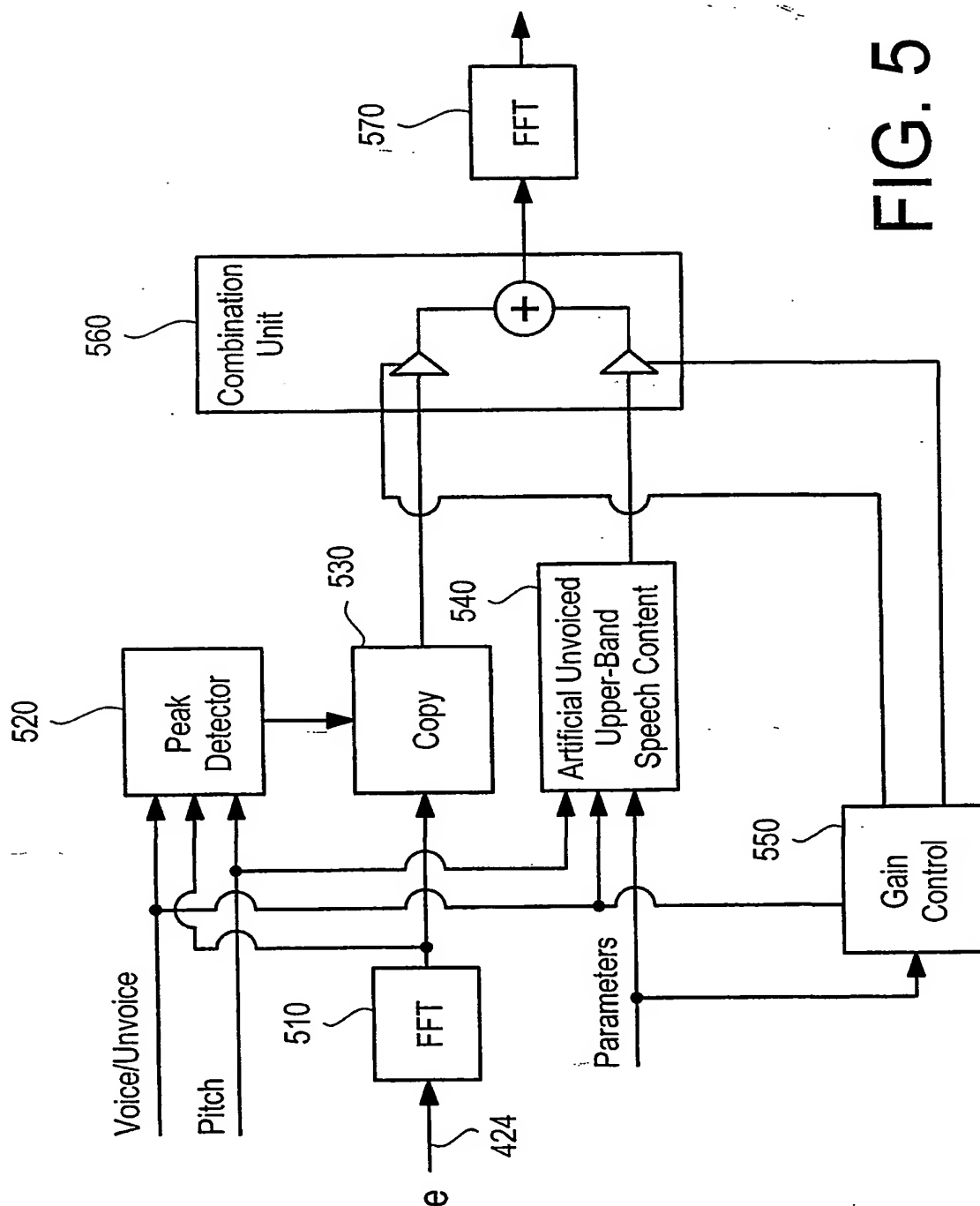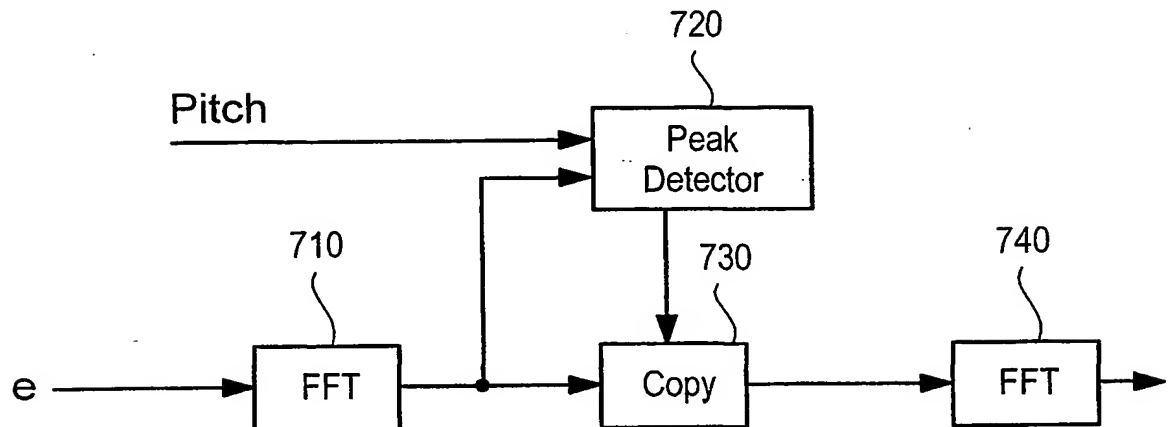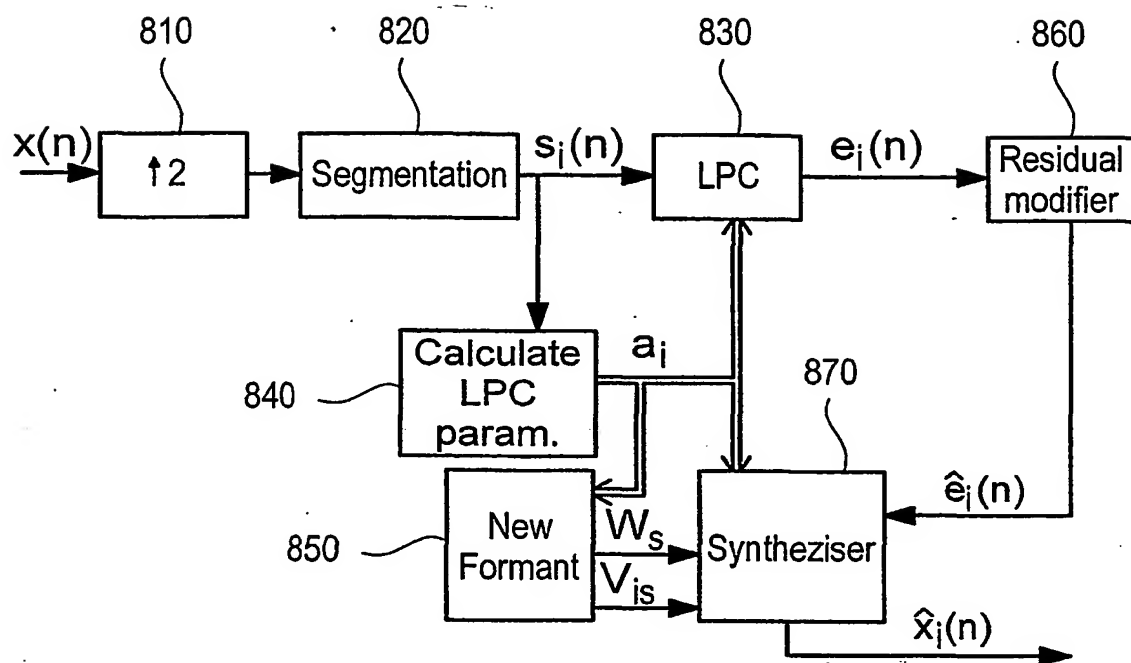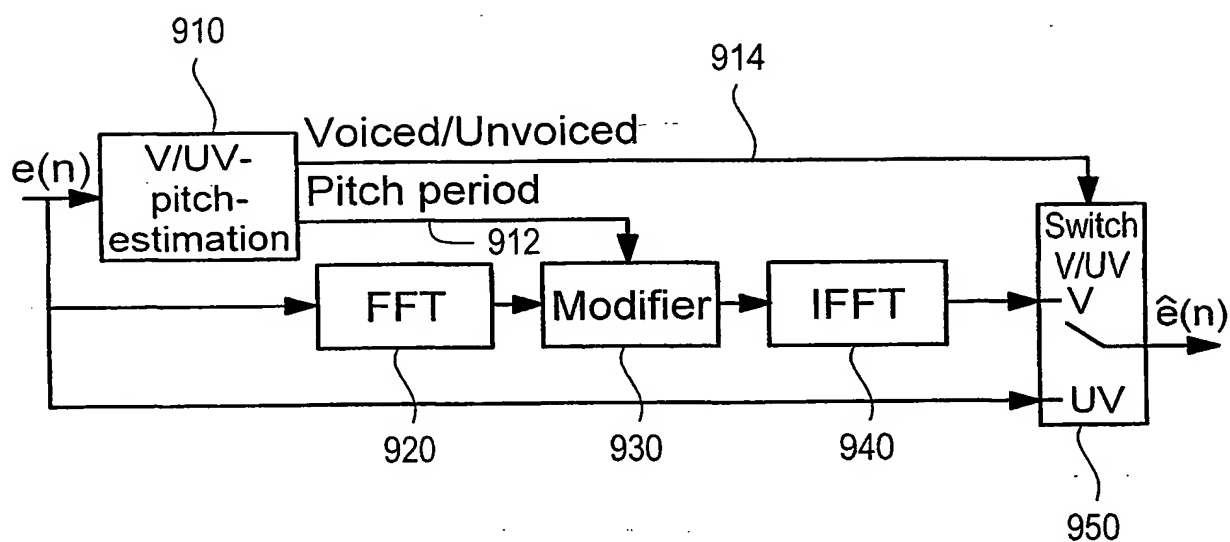
110

Receiver	Channel decoder	Vocoder	Bandwidth extention	D/A-converter	Speaker

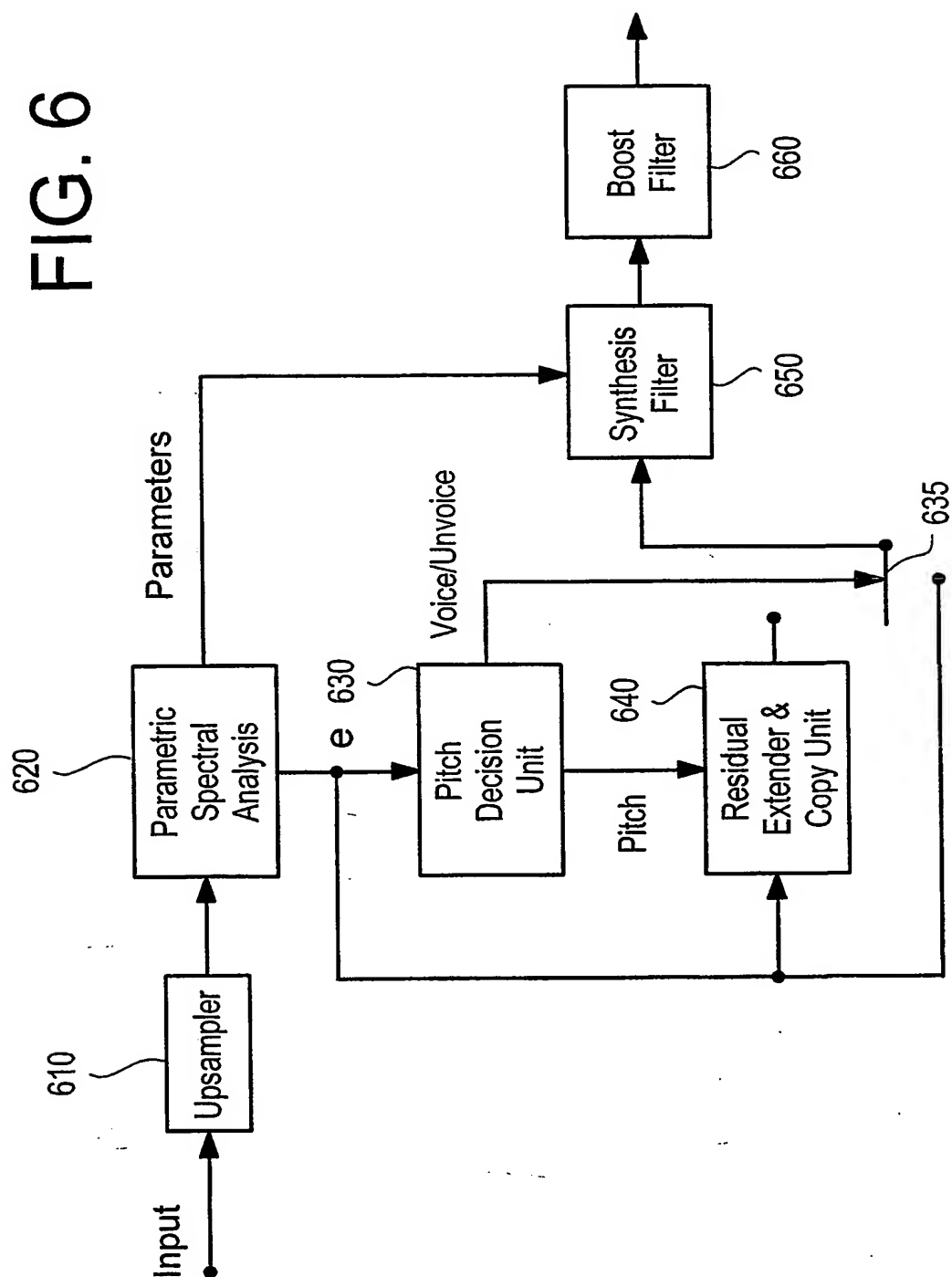120 130 140 150 160 170

FIG. 2

FIG. 3

FIG. 4

BNSDOCID: <WO   0156021A1_L_>

5/15

FIG. 5

FIG. 6

FIG. 7

8/15

FIG. 8

FIG. 9

FIG. 10

FIG. 11

FIG. 12

$$x(n) \longrightarrow \boxed{H(z)} \longrightarrow y(n)$$

$$x(n) \longrightarrow \boxed{H_{i1}(z)} \longrightarrow \boxed{H_{i2}(z)} \longrightarrow y(n)$$
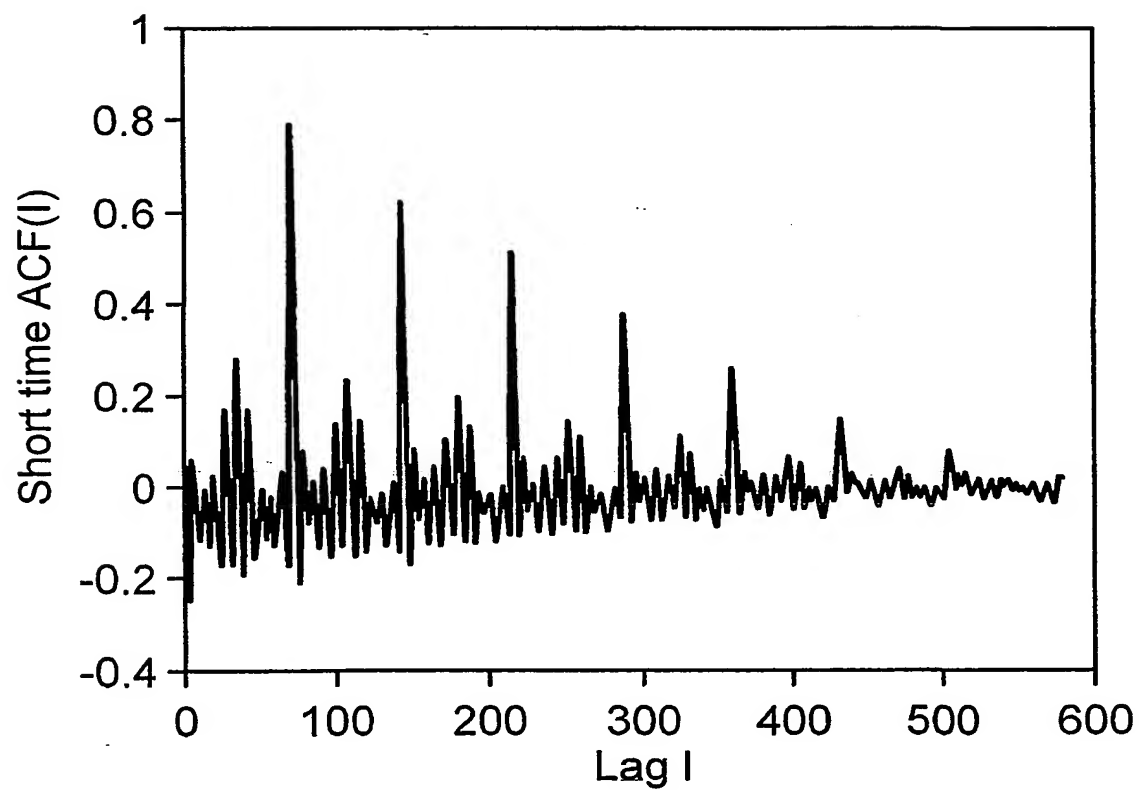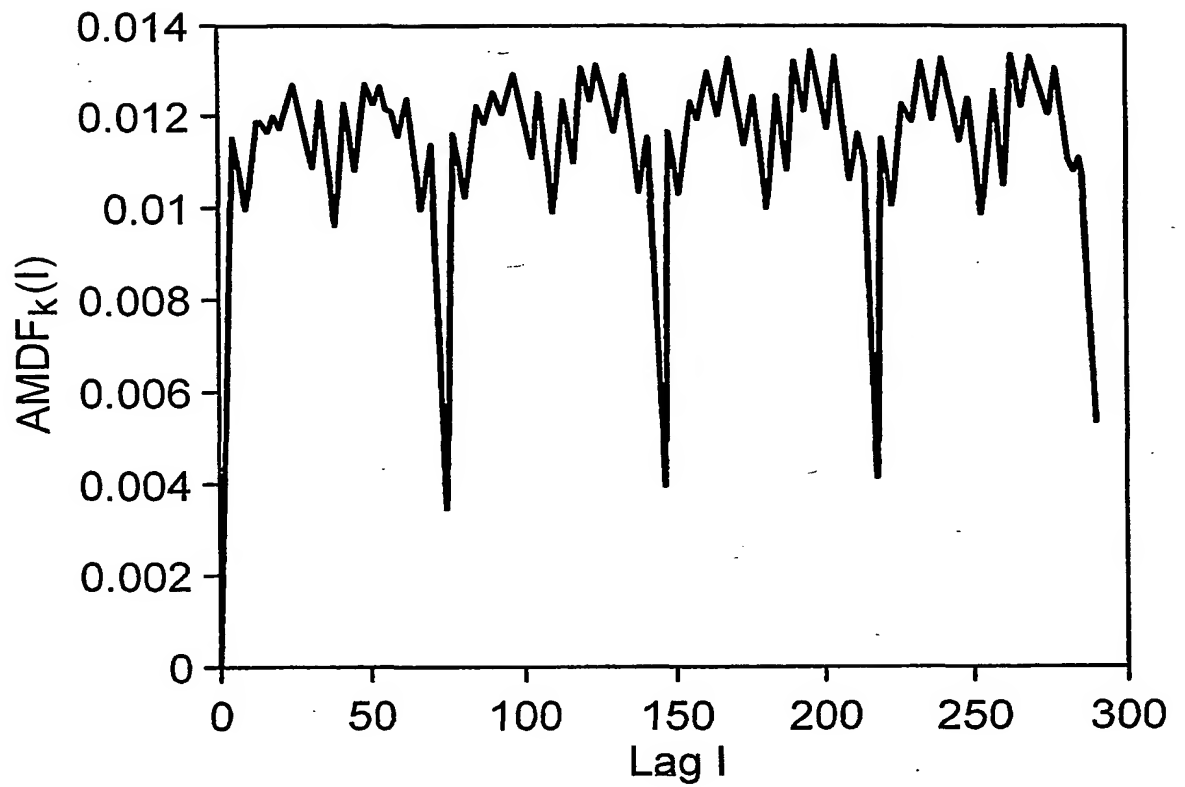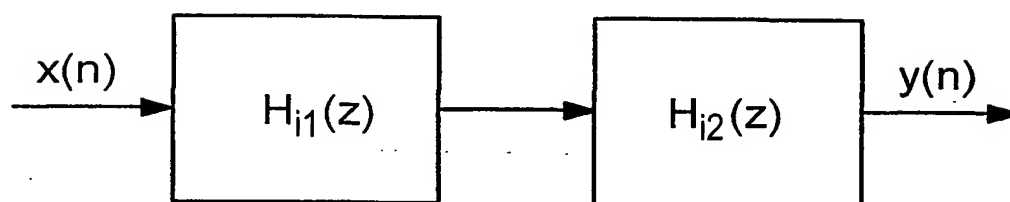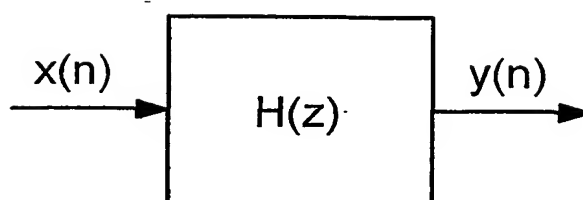
FIG. 13
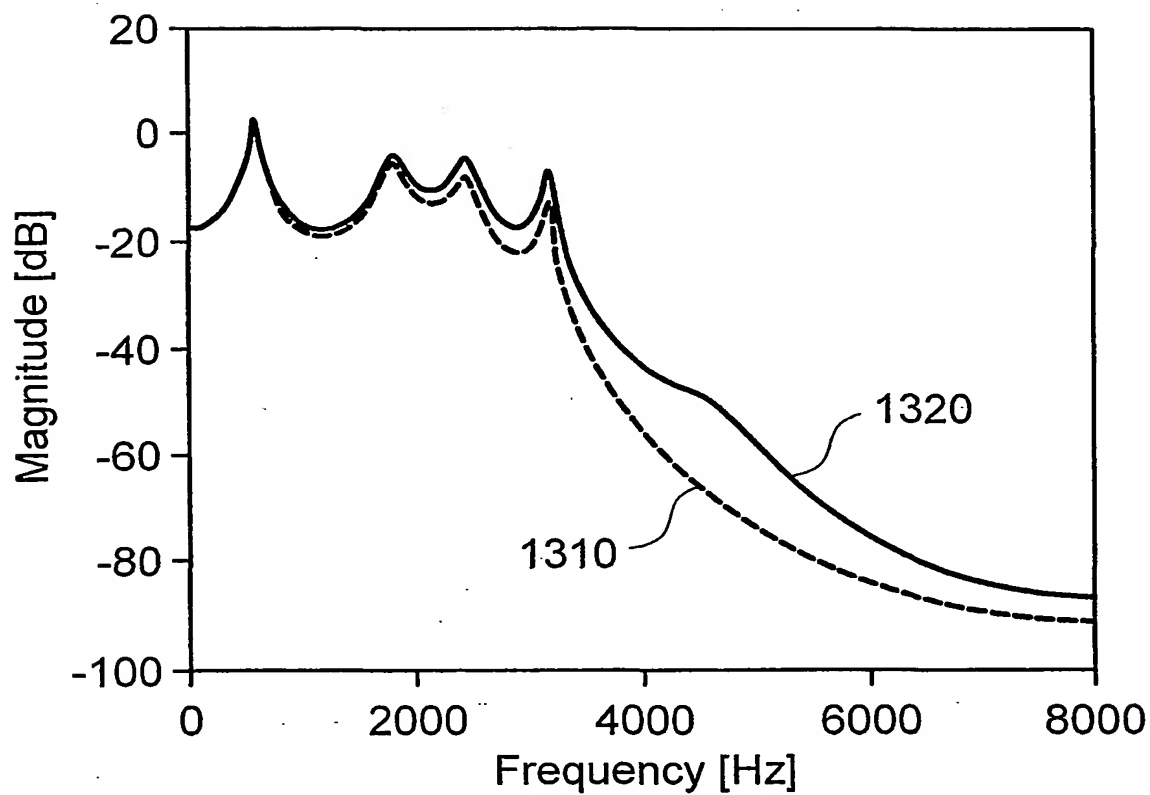
FIG. 14
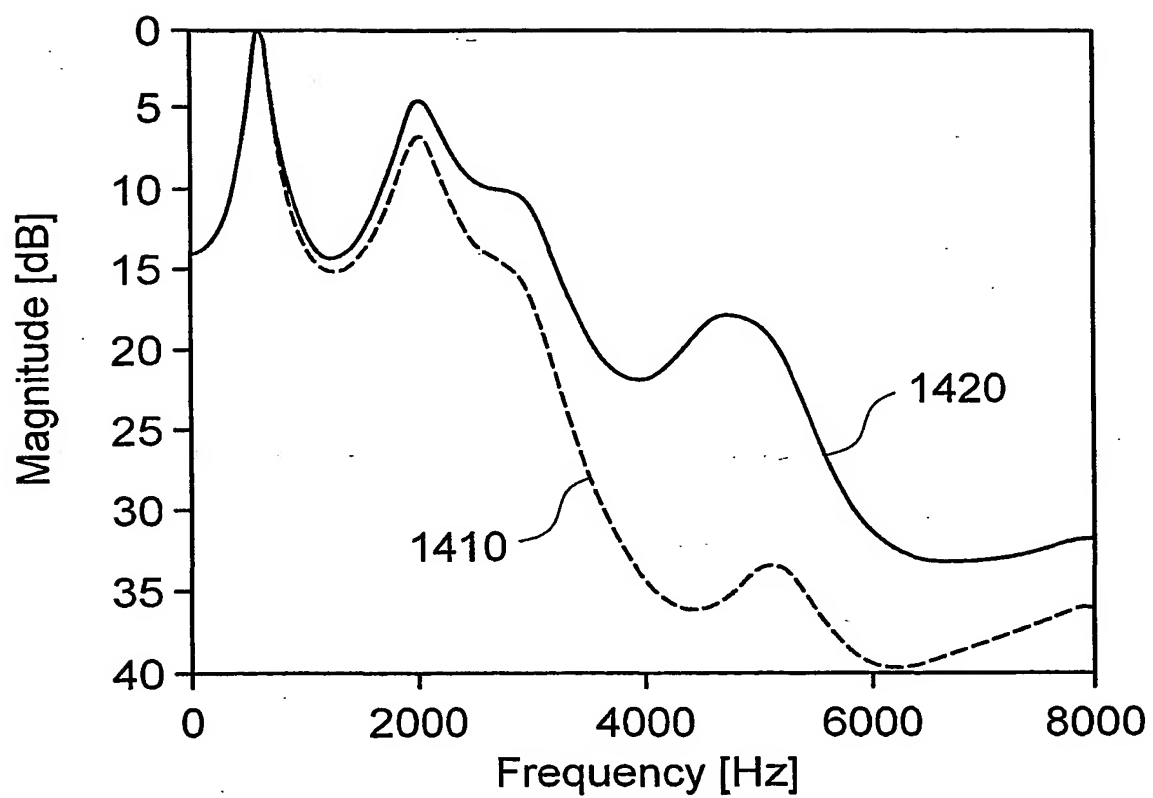
FIG. 15
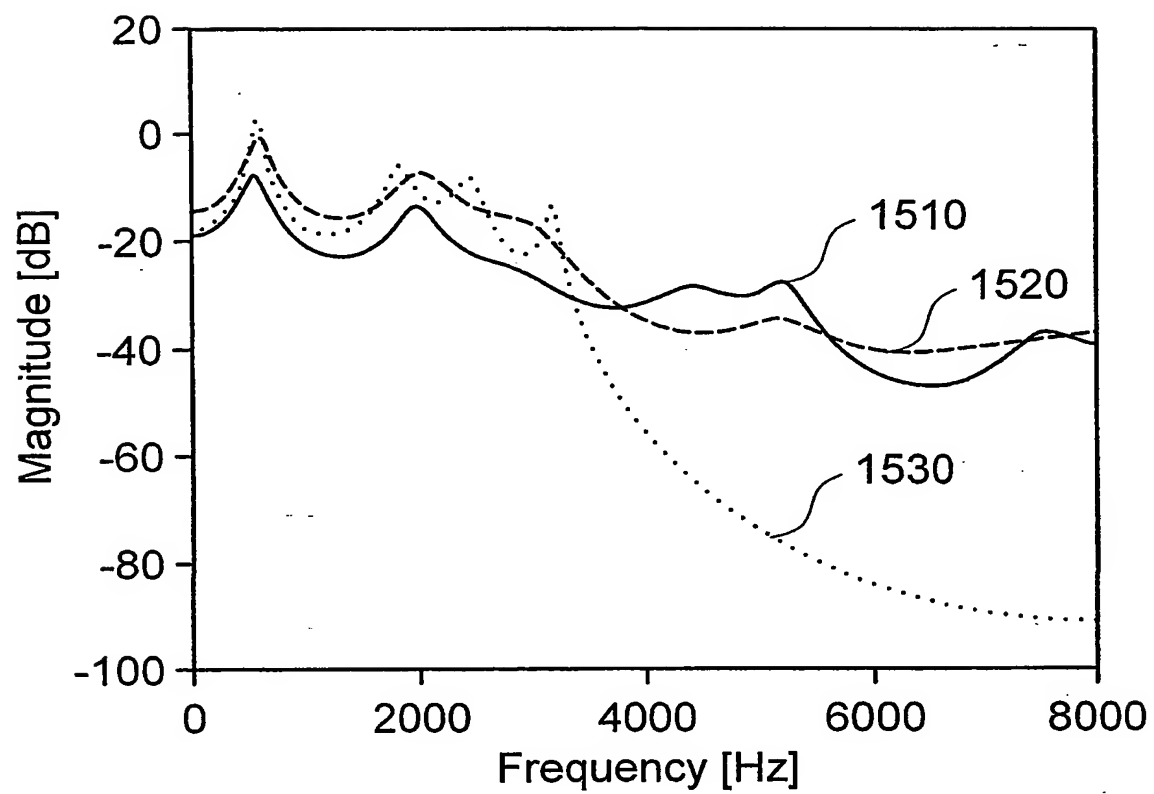
INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L21/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

INSPEC, IBM-TDB, EPO-Internal, PAJ, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	YAN MING CHENG ET AL: "Statistical recovery of wideband speech from narrowband speech" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, IEEE INC. NEW YORK, US, vol. 2, no. 4, October 1994 (1994-10), pages 544-548, XP002106825 ISSN: 1063-6676 abstract	1,9
A	paragraph '0003!	3,10
A	EP 0 945 852 A (BRITISH TELECOMM) 29 September 1999 (1999-09-29) abstract	3,10,16, 17

☐ Further documents are listed in the continuation of box C.   ☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 May 2001	21/05/2001

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Van Doremalen, J

Form PCT/ISA/210 (second sheet) (July 1992)

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
EP 0945852	A	29-09-1999	EP 1064648 A		03-01-2001
			WO 9949454 A		30-09-1999